News
PENNSTATE

(814) 863-4682 FAX (814) 863-2246

Office of Media Relations and Public Information
Eberly College of Science

520 Thomas Building
The Pennsylvania State University
University Park, PA 16802

EMBARGOED: Publication of information about the research described here is prohibited -- in any medium -- by the journals including *Nature* until 1:00 p.m., U. S. Eastern time, on 5 September 2012.

Huge Human Gene Study Includes Penn State Research

The first integrated understanding of how the human genome functions will be published this week -- the triumphant result of a collaborative five-year project involving more than 440 researchers working in 32 labs worldwide. The Encyclopedia of DNA Elements project, known as ENCODE, will publish simultaneously on 6 September 2012 a massive number of scientific papers, including 1 main integrative paper and 5 others in *Nature*; 18 in *Genome Research*; 6 in *Genome Biology*; and other affiliated papers in *Science*, *Cell*, and other scientific journals.

During the ENCODE study, researchers found that more than 80 percent of the human genome sequence is linked to biological function. They also mapped more than 4 million regulatory regions where proteins interact with the DNA with exquisite specificity. These findings are a significant advance in understanding the precise and complex controls over the expression of genetic information within a cell.

"Penn State's contribution to the ENCODE project involves using the new ENCODE data to help explain how genetic variants that do not affect the structure of encoded proteins could affect a person's susceptibility to disease," said Ross Hardison, the T. Ming Chu Professor of Biochemistry and Molecular Biology at Penn State and a member of the ENCODE research team. The research led by Hardison is highlighted in the main integrative ENCODE paper to be published in the journal *Nature*.

"Genome-wide association studies can map with high resolution the places on our genomes where variation in the DNA sequence among individual persons affects their likelihood of having diabetes, cardiac disease, any of a large number of autoimmune diseases such as Crohn's disease, and other common diseases," Hardison said. Because most of these genetic variations are not in regions of the DNA that contain the codes for producing proteins, scientists suspected that some of these non-coding regions might have an important role in controlling the expression of genes.

Hardison's team at Penn State worked with others in the ENCODE Consortium to show, on a genome-wide scale, that many of the DNA regions that do not hold codes for proteins do, indeed, have an important role in controlling which genes are turned on and which are turned off. "Moreover, our research has made it possible to generate specific molecular hypotheses for how genetic variants in these DNA regions that control gene expression could affect the susceptibility to disease," Hardison said. "We demonstrate this process using, as an example, a locus associated with Crohn's and a few other autoimmune diseases. It is exciting to see our basic research revealing insights that help the progress of medical science, potentially facilitating a more personalized approach to medical practice."

In addition to Hardison, other Penn State scientists whose work on the ENCODE project is featured among the papers to be published on 6 September include Programmer/Analyst Belinda Giardine, Postdoctoral Scholars Robert S. Harris and Weisheng Wu, and Professor of Biology and of Computer Science and Engineering Webb Miller.

The overall ENCODE findings bring into much sharper focus the continually active genome in which proteins routinely turn genes on and off using sites that are sometimes at great distances from the genes they regulate; where sites on a chromosome interact with each other, also sometimes at great distances; where chemical modifications of DNA influence gene expression; and where various functional forms of RNA, a form of nucleic acid related to DNA, help regulate the whole system. "The ENCODE catalog is like Google Maps for the human genome," said Elise Feingold, a program director at the National Institutes of Health National Human Genome Research Institute (NHGRI), who helped to start the ENCODE Project. "The ENCODE maps allow researchers to inspect the chromosomes, genes, functional elements and individual nucleotides in the human genome in much the same way."

"During the early debates about the Human Genome Project, researchers had calculated that only a few percent of the sequence encoded proteins, the workhorses of the cell," said Eric D. Green, director of NHGRI. "Early on, some scientists even argued that most of the genome was 'junk.' ENCODE now gives us much more appreciation of the complex molecular ballet that converts genetic information into living cells and organisms, and we can now say that there is very little, if any, junk DNA."

Hundreds of researchers in the United States, United Kingdom, Spain, Singapore, and Japan performed more than 1,600 sets of experiments on 147 types of tissue with technologies standardized across the consortium. The experiments relied on innovative uses of new next-generation sequencing technologies enabled, in part, by NHGRI's technology initiative for DNA sequencing. In total, ENCODE generated more than 15-trillion bytes of raw data and its analysis consumed the equivalent of more than 300 years of computer time.

The ENCODE project received principal funding from the National Human Genome Research Institute at the National Institutes of Health. Computation was enabled, in part, through the Penn State Cyberstar Computer, funded by the National Science Foundation (grant OCI-0821527).

[Barbara K. Kennedy]

CONTACTS
Ross Hardison: rch8@psu.edu, +1 814-863-0113
Barbara Kennedy (PIO): science@psu.edu, +1 814-863-4682

IMAGE
A high-resolution image associated with this research is online at http://science.psu.edu/news-and-events/2012-news/Hardison9-2012

IMAGE CAPTION
ENCODE is a massive database cataloging many of the functional elements of the entire collection of human genes -- the human genome. The ENCODE data are being made available to the scientific community and to the public as an open resource. This illustration shows a group of proteins bound to the spiraling strands of DNA, which is the genetic material. A chromosome, composed of tightly coiled DNA, is illustrated in the background. The new papers from the ENCODE consortium show that differences in the way certain proteins interact with DNA in each person play a role in their susceptibility to some common diseases.

MORE INFORMATION
The ENCODE Consortium placed the resulting data sets as soon as they were verified for accuracy, prior to publication, in several databases that can be freely accessed by anyone on the Internet. These data sets, as well as more information about ENCODE, are available at the ENCODE project portal http://www.encodeproject.org.

"Because the ENCODE project has generated so much data, we, together with the ENCODE Consortium, have introduced a new way to enable researchers to navigate through the data," said Magdalena Skipper, senior editor at *Nature*, which produced the freely available publishing platform on the Internet. Since the same topics were addressed in different ways in different papers, the new website, http://www.nature.com/encode/, will allow anyone to follow a topic through all of the papers in the ENCODE publication set in which it appears, by clicking on the relevant "thread" at the Nature ENCODE explorer page. ENCODE scientists believe this tool will illuminate many biological themes emerging from the analyses.